Comprehensive targeted super-deep next generation sequencing enhances differential diagnosis of solitary pulmonary nodules

Ye, Mingzhi; Li, Shiyong; Huang, Weizhe; Wang, Chunli; Liu, Liping; Liu, Jun; Liu, Jilong; Pan, Hui; Deng, Qiuhua; Tang, Hailing; Jiang, Long; Huang, Weizhe; Chen, Xi; Shao, Di; Peng, Zhiyu; Wu, Renhua; Zhong, Jing; Wang, Zhe; Zhang, Xiaoping; Kristiansen, Karsten; Wang, Jian; Yin, Ye; Mao, Mao; He, Jianxing; Liang, Wenhua

Published in:
Journal of Thoracic Disease

DOI:
10.21037/jtd.2018.04.09

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Original Article

Comprehensive targeted super-deep next generation sequencing enhances differential diagnosis of solitary pulmonary nodules

Mingzhi Ye[1,2,3,4,5]*, Shiyong Li[1,4]*, Weizhe Huang[2]*, Chunli Wang[6,7]*, Liping Liu[2], Jun Liu[6,7], Jilong Liu[1], Hui Pan[2], Qiuhua Deng[2], Hailing Tang[2], Long Jiang[2], Weizhe Huang[2], Xi Chen[6,7], Di Shao[1], Zhiyu Peng[4], Renhua Wu[6,7], Jing Zhong[4], Zhe Wang[4], Xiaoping Zhang[4], Karsten Kristiansen[5], Jian Wang[4], Ye Yin[4], Mao Mao[4], Jianxing He[2], Wenhua Liang[2]

[1]BGI-Guangzhou Medical Laboratory, BGI-Shenzhen, Guangzhou 510006, China; [2]The First Affiliated Hospital of Guangzhou Medical University, National Clinical Research Center for Respiratory Disease, State Key Laboratory of Respiratory Disease, Guangzhou 510120, China; [3]BGI-Guangzhou, Guangzhou Key Laboratory of Cancer Trans-Omics Research, Guangzhou 510006, China; [4]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China; [5]Department of Biology, University of Copenhagen, Copenhagen, Denmark; [6]Tianjin Medical Laboratory, [7]Binhai Genomics Institute, BGI-Tianjin, BGI-Shenzhen, Tianjin 300308, China

Contributions: (I) Conception and design: J He, M Mao; (II) Administrative support: Z Peng, X Zhang, Y Yin, W Wang; (III) Provision of study materials or patients: W Liang, L Liu, H Pan; (IV) Collection and assembly of data: Q Deng, L Jiang, W Huang; (V) Data analysis and interpretation: S Li, C Wang, J Liu, X Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to the work.

Correspondence to: Wenhua Liang, MD; Jianxing He, MD, PHD. Department of Thoracic Surgery and Oncology, The First Affiliated Hospital of Guangzhou Medical University, National Clinical Research Center for Respiratory Disease, State Key Laboratory of Respiratory Disease, No. 151 Yanjiang Rd., Guangzhou 510120, China. Email: liangwh1987@163.com; drjianxing.he@gmail.com.

Background: A non-invasive method to predict the malignancy of surgery-candidate solitary pulmonary nodules (SPN) is urgently needed.

Methods: Super-depth next generation sequencing (NGS) of 35 paired tissues and plasma DNA was performed as an attempt to develop an early diagnosis approach.

Results: Only ~6% of malignant nodule patients had driver mutations in the circulating tumour DNA (ctDNA) with >10,000-fold sequencing depth, and the concordance of mutation between tDNA and ctDNA was 3.9%. The first innovative whole mutation scored model in this study predicted 33.3% of malignant SPN with 100% specificity.

Conclusions: These results showed that lung cancer gene-targeted deep capture sequencing is not efficient enough to achieve ideal sensitivity by simply increasing the sequencing depth of ctDNA from early candidates. The sequencing could not be evaluated hotspot mutations in the early tumour stage. Nevertheless, a larger cohort is required to optimize this model, and more techniques may be incorporated to benefit the SPN high-risk population.

Keywords: Solid pulmonary nodule; early diagnosis; circulating tumour DNA (ctDNA); lung cancer; tumor mutational burden (TMB)

Submitted Mar 15, 2018. Accepted for publication Mar 26, 2018.
doi: 10.21037/jtd.2018.04.09
View this article at: http://dx.doi.org/10.21037/jtd.2018.04.09

Introduction

Lung cancer continues to be the leading cause of cancer mortality in both men and women worldwide (1). Early diagnosis is crucial for improving lung cancer survival, given that the prognosis of stage I lung cancer is considerably favourable with a 5-year survival rate of more than 70% compared with metastatic late-stage disease (<5% survival) (2). Currently, the most successful method for early detection is low-dose computed tomography (LDCT) scan screening, which was demonstrated by the National Lung Cancer Screening Trial (NLST) study to reduce mortality by 20% compared with chest radiograph screening of lung cancer (3).

The widespread application of LDCT has led to a significant increase in the detection of lung nodules (4,5). The prevalence of solitary pulmonary nodules (SPN) (<3 cm in diameter) is 10–20% in the United States (6) and is higher in people with Asian ancestry probably due to genetic and environmental factors. Most SPN found in CT scans are benign, even among high-risk populations such as smokers. A few algorithms or prediction models based on nodule features in the CT scan have been developed; however, their accuracy remains unsatisfactory (7). On one hand, timely identification of malignant nodules is crucial because they represent a localized disease and are potentially curable. On the other hand, it is costly and possibly harmful to manage an SPN with radiation exposure from repeated CT scans or invasive procedures such as biopsy or surgical resection that are associated with potential morbidity and induce unnecessary anxiety. Therefore, there is a critical need for additional tests that can further stratify the SPN found by LDCT as malignant and non-malignant.

Non-invasive tests are preferable. ^{18}F-FDG-PET/CT only slightly adds to diagnostic value, and its use is limited by its low cost-effectiveness (8). A few plasma biomarkers, such as CEA and CA-125, have been used to screen and diagnose lung cancers (9-11). However, the sensitivity of serum biomarkers is relatively low because they are proteins and thereby will be elevated only when the tumour burden is high. Therefore, there is no sufficiently reliable biomarker that exhibits both high sensitivity and specificity for the diagnosis of malignant SPN. ctDNA represents a promising option: it is released or excreted by tumour cells, circulates in the blood of a patient with cancer, and can serve as direct evidence of malignancy (12).

Because of the diverse mutation pattern of lung cancer, it cannot be evaluated using conservative single-gene mutations or hotspot mutations. Unlike PCR-based techniques, NGS simultaneously allows the detection of a wide spectrum of loci. Comprehensive analyses could theoretically increase sensitivity. In addition, genetic mutations should be more reliable than other qualitative markers (e.g., antibody or micro-RNA level), which require tricky cut-offs.

Previously, a report described using the total plasma cell-free DNA (cfDNA) level to discriminate non-small cell lung cancer (NSCLC) from benign lung pathologies and healthy controls with 86.4% sensitivity and 61.4% specificity (13). However, debates remain regarding the lower limit of detection of ctDNA NGS. We hypothesized that the analysis of the lung cancer-related somatic mutations from ctDNA could provide better opportunities for minimally invasive SPN diagnosis. We hereby aimed to develop a practical tool based on ctDNA profiling and super-deep sequencing methods and test its ability to distinguish between malignant and non-malignant SPN in this pilot study. However, debates remain regarding the lower limit of detection of ctDNA NGS.

Results

Lung peripheral nodule clinical features and tumour serum protein marker classification

A total of 1,254 consecutive candidate patients were reviewed for resection of lung peripheral nodules in the First Affiliated Hospital of Guangzhou Medical University. In postoperative pathological examination, 69% of lung peripheral nodules were diagnosed as malignant, and the distribution of malignant SPN subtypes is shown in *Figure S1*. Almost 80% of SPNs were adenocarcinoma, which included 13% AAH (atypical adenomatous hyperplasia)/AIS (adenocarcinoma in situ)/MIA (minimally invasive adenocarcinoma) malignancy patients. Surgery and biopsy were risky for the patients and could cause complications. A non-invasive method was required to identify the malignancy of surgery-candidate lung peripheral nodules. Tumour serum protein markers (CEA, NSE, CA125, CA153, and CYFRA21-1) are conventionally used to determine the malignancy of SPN. However, only CEA and CYFRA21-1 in malignant SPN were significantly higher than in non-malignant SPN. The expression of NSE and CA153 in this statistical cohort was not different between malignant and benign cases based on the p-value calculated by the unpaired *t*-test (*Figure 1*). The expression of CA125 in benign SPN was significantly higher than in malignant SPN. The mean expression of

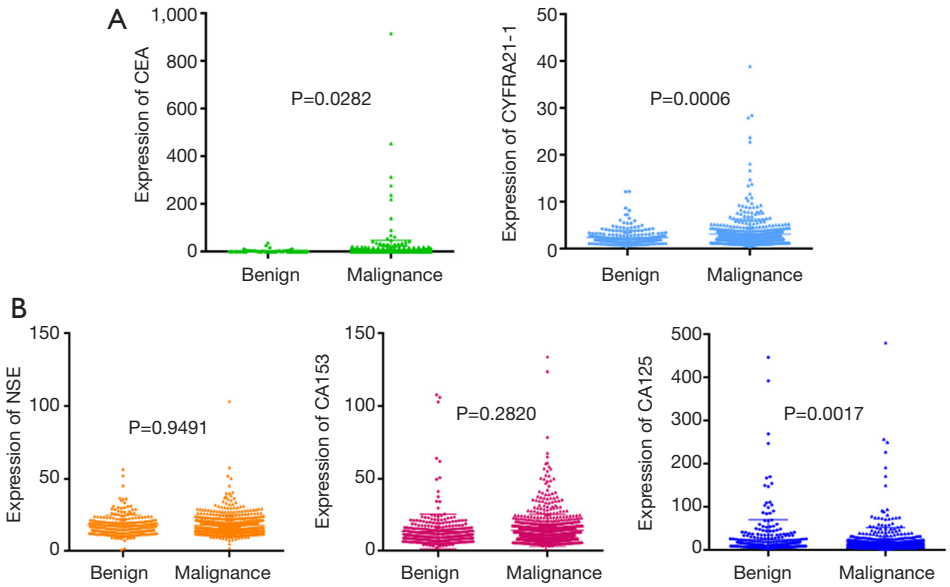

Figure 1 Clinical distribution of all SPNs. The pie chart (A) shows the distribution of SPN histological types. The expression comparison of tumour serum protein markers is shown in (B). P values were calculated by an unpaired *t*-test. Different colour indicates different markers; dot indicates benign SPN; little triangle indicates malignant SPN. Other cancer, malignant SPNs that cannot be categorized by the listed cancer types. SCLC, small cell lung cancer; SPN, solitary pulmonary nodule.

serum protein markers was similar. Therefore, this signature limits serological indicators as an accurate diagnosis of early lung cancer. Neither CEA (cut-off 5 ng/mL, specificity 90.1%, sensitivity 23.8%) or CYFRA21-1 (cut-off 3.3 ng/mL, specificity 80.6%, sensitivity 28.5%) nor their combination (specificity 77.6%, sensitivity 42.1%) could precisely predict malignancy. When using 10 ng/mL as the cut-off, CEA achieved a specificity of 97.1%, but the sensitivity was only 9.7%.

ctDNA seemed to be a good option as it has been largely reported and named as a non-invasive method for patients' targeted genes tests and recurrent monitoring (14,15). In this study, both surgically resected lung peripheral nodules and plasma DNA were investigated by extra-deep high throughput sequencing of at least 10,000-fold depth to classify malignancies or non-malignancies in the early stage of lung cancer. The pipeline of this research is shown *Figure S2*. Thirty-five prospective samples were consecutively collected to perform the next generation sequencing (NGS) to develop a non-invasive malignant peripheral nodule prediction method. All the lung surgery candidate nodules, formalin-fixed, paraffin embedded (FFPE) tissues, and corresponding blood samples were collected as controls. Clinical summary information of the selected patients with lung pulmonary nodules is shown in *Table 1*; 62.9% of the patients were males, and 37.1% were females. Clinical histological results identified malignant peripheral nodules in 31 out of 35 patients and benign peripheral nodules in the remaining 4 patients. Out of 31 malignant SPNs, 81% of patients were diagnosed with lung adenocarcinoma, which had a much higher distribution than our statistical cohort, likely because of the relatively small sample size. All included cases were in clinical stage I. In postoperative pathological evaluation, 83.9% of the patients remained in stage I, while 29% of patients had advanced to stage II and III which were diagnosed postoperatively by incidental finding of positive lymph nodes. Therefore, all cases are necessary to be screened by NGS to explore the genomic profile. Each sample's detailed clinical information is recorded in Supplementary *Table S1*.

Landscape of somatic mutations and driver genes

DNA from white blood cells was used as a corresponding normal control to detect somatic mutations from FFPE and ctDNA samples. All of the samples were analysed by lung cancer target capture and sequenced by the Illumina HiSeq 2000 instrument. The lung cancer panel included the exon region of lung cancer driver genes and top mutational lung adenocarcinoma-related genes, based on the COSMIC

Table 1 Clinical information of patients with lung pulmonary nodules sampled for ultra-deep sequencing

Clinical feature (35 total samples)	Data
Age at surgery [median, range]	59 [27–81]
Gender	
Male	22
Female	13
Cancer type	
Benign	4
Cancer	31
AIS/MIA	2
Adenocarcinoma	25
Squamous carcinoma	1
Other cancer	3
Clinical tumour stage	
I	35
Pathological tumour stage	
I	21
II*	5
III#	4
NA	1

*, unexpected N1 lymph nodes incidentally found by pathological examination; #, unexpected N2 lymph nodes incidentally found by pathological examination. AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma.

database (*Table S2*). Followed by deep sequencing, at least 99.9% of the target genomic regions of each case were covered (*Table S3*). The median depths were 600× (from 171 to 1,941) for the 38 FFPE samples, 823× (from 524× to 2,543×) for the 38 normal control samples, and 1,896× (from 610 to 7,653) for the 35 ctDNA samples.

In total, 89 non-silent SNV/InDels/SV (range from 0 to 11) were discovered in the 31 tumour tissue samples. No mutations were detected from the 4 benign pulmonary nodules (*Figure 2*). Twenty-nine of the 31 cancer samples contained at least one non-silent mutation, and non-silent mutations were detected in all of the lung adenocarcinomas. The two samples in which mutations were not detected were lung carcinoid tumours. This might be because of the limitation of the panel's gene list, which was based on lung adenocarcinoma and was not available for lung carcinoid tumours and squamous carcinoma. Each sample's detailed mutational information, excluding that of SV, is shown in *Table S4*. Four fusions were detected in 3 samples, and 2 of them (*ALK*, *ROS1*) were found in sample S23_A. *ALK* was found in sample S5_A, and RET was found in sample S27_A. However, only 3 out of 4 were successfully validated by immunohistochemistry, except *ROS1* fusion in sample S23 (*Table S5*). Twenty-eight non-silent mutations (SNV/InDel) were detected in the corresponding plasma samples. Only 1 non-silent mutation was detected in the plasma of benign sample S32_N (*Figure 2*, *Table S6*). Clinical information of each patient is shown in *Figure 2*, with the distribution of stage, subtypes, sample types, and gender. The malignant SPN were divided into two groups according to the tumour stage in *Figure 2* (stage I *vs.* stage II-III). For the patients with SPN, each patient's mutational number had no difference in the tissue or plasma samples (*Figures 2,S3*). Compared with the mutations in the tissue of benign cases, mutations detected in the tumour tissue had a significantly higher mutational ratio. The mutational number from ctDNA was also assessed with respect to the size of SPN or the mutations in tissue, but no correlation was found (*Figures 2,S4*). After comparison of the mutational consistencies between tissue and plasma (*Figure S5*), only 6 out of 152 mutations detected in FFPE were found in the corresponding ctDNA samples; the concordance between ctDNA and FFPE samples was much lesser than that in a previously reported study (16). Even more, 5 out of the 6 overlapping mutations came from one sample (S8_A), which shows a lack of efficacy in early stage ctDNA evaluation to some degree.

Well-known driver genes were detected in 22 out of 31 (71%) malignant FFPE samples, and the frequency of each driver gene was as follows: *EGFR*: 46%, *KRAS*: 3%, *ALK/ROS1/RET* fusion: 11%, *BRAF*: 3%. The frequency of each driver gene was different from that in our previous study (17,18), especially *KRAS*, which might be because of sample size limitation and the sequencing panel, which was designed only for lung adenocarcinoma. Except for sample S23_A with non-validation *ROS1* fusion, all the other SPNs had a unique driver gene (*Figure 3*). Of the *EGFR*-positive SPN samples, 37.5% had compound *EGFR* mutations; 8 samples contained L858R mutations, and 7 samples had exon 19 deletions (*Table 2*). Even 3 *EGFR* mutations were found in sample S22_A. Although all the *EGFR* compound mutations were rare, SNV/InDel and the co-*EGFR* mutational ratio were higher than in a recent Asian study (19), This previous study proved that patients with a single *EGFR* mutation had better survival rates than patients with compound *EGFR*

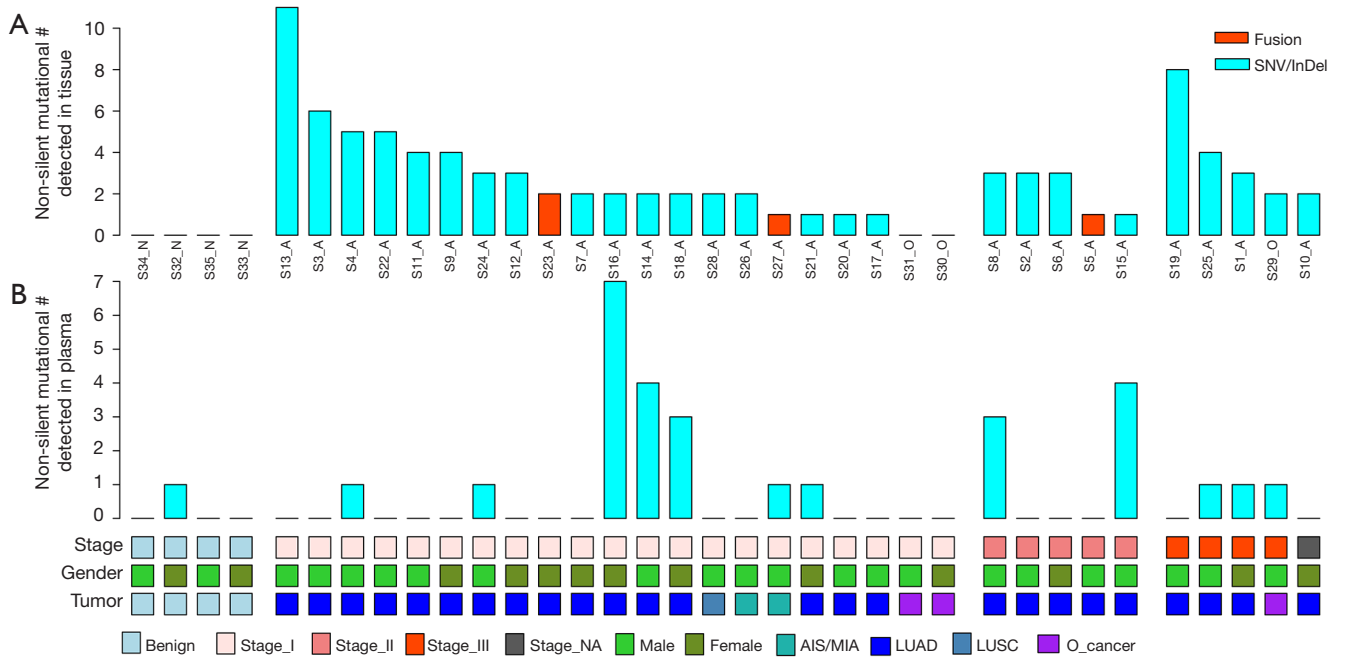

Figure 2 Somatic mutation landscape of FFPE and ctDNA samples. (A) The bars represent the non-silent mutational number of each sample. The samples are sorted by the tumour stage [benign, stage I (AIS/MIA, adenocarcinoma), and stage II–III] and number of non-silent mutations. Mutational type is distinguished by colour. ctDNA mutational landscape is shown in (B). The major clinical information (plasma, gender, tumour stage, cancer subtype) is shown in (C). ctDNA, circulating tumour DNA; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; FFPE, formalin-fixed, paraffin embedded.

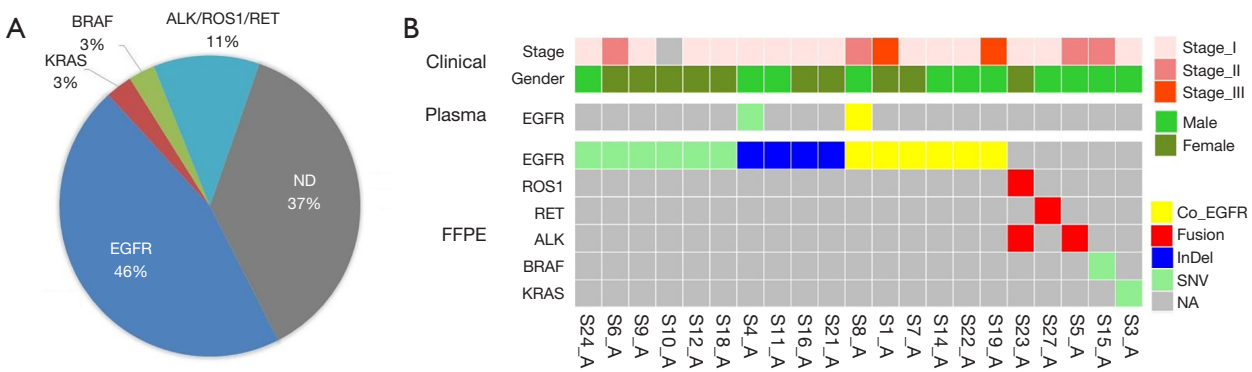

Figure 3 Driver gene mutations detected in paired samples. Distribution of driver gene mutational frequency is shown in (A), and each patient's driver gene in FFPE and plasma samples is shown in (B). FFPE, formalin-fixed, paraffin embedded.

mutations, and there were no differences in disease-free survival rates. All the *EGFR* co-mutation samples had similar allele frequencies of each *EGFR* mutation. Driver mutations were found in only two ctDNA samples, both of which were mutated in *EGFR*, but the *EGFR* driver mutation in sample S4 was different in the tumour tissue and plasma. An *EGFR* compound mutation (L858R + S768I) was detected both in sample S8_A tissue and plasma samples. Overall, the extremely low driver mutation concordance between the FFPE and corresponding ctDNA suggested that ctDNA content in SPN patients was also too low to be efficiently sequenced by the NGS method. ddPCR was performed as a sensitive tool for low-frequency mutation testing to validate known hotspot driver mutations detected in the FFPE and

ctDNA samples.

Driver mutation validation by ddPCR

ddPCR is a well-known low-frequency mutation detection platform and serves as an efficient tool to test the reliability of sequencing data from NGS. As for the limitation of ctDNA quantity, only 6 samples with *EGFR/KRAS* hotspot driver mutations were validated to confirm the mutation accuracy and frequency detected by NGS (*Table 3*). Meanwhile, four corresponding FFPE samples were also randomly selected to be validated by ddPCR as a control. A similar mutant allele frequency of FFPE samples was observed with NGS and ddPCR. Results from ddPCR detection indicated that there was a good concordance between the NGS and ddPCR detection, providing favourable evidence that the sequencing data are reliable. The ddPCR results helped prove that the mutational concordance between ctDNA and FFPE of SPN was much lower than in a previous study (16).

Table 2 *EGFR* mutational landscape

EGFR mutational type	N=16	%
L858R	5	31.2
Exon 19 deletion	4	25.0
Rare SNV	1	6.3
Co-mutation	6	37.5
L858R + R889G	1	6.3
L858R + S768I	1	6.3
L858R + V834L	1	6.3
Exon19del + L833V	1	6.3
Exon19del +A750Pro	1	6.3
Exon19del + N756H+ A755G	1	6.3

Malignant lung peripheral nodule prediction

A new model used to predict the malignancy of lung peripheral nodules based on the 35 plasma samples was developed using two matrixes: (I) the score of mutation (MS) contributing to the lung adenocarcinoma genesis and development and (II) the tumour burden of cfDNA (TMB), which was used to evaluate the whole mutational frequency within the panel region. Cancer gene census from the COSMIC database was used to divide the mutational genes into three groups: (I) oncogene, (II) tumour suppressor gene, and (III) non-cancer-related gene. Well-known LUAD driver mutation was added as a fourth group (such as *EGFR*:L858R, *KRAS*:G12V, *ROS1/RET/ALK* fusion, and so on). The score of each mutation was assigned based on the formula below:

$$MS = sum(S_i)\ S_i = \begin{cases} 2^3\ class1: well\text{-}know\ LUAD\ driver\ mutation \\ 2^2\ class2: oncogenic\ mutation\ except\ class1 \\ 2^1\ class3: mutation\ in\ tumor\ suppressor\ gene \\ 2^0\ class4: mutation\ except\ before\ class \end{cases}$$

i = *mutation in one sample*

Table 3 Mutational frequency detected by NGS and ddPCR

Sample ID	Gene	Mutation	Sample type	NGS freq	ddPCR freq
S8_A	EGFR	p.L858R	ctDNA	0.0094	0.0049
S8_A	EGFR	p.L858R	FFPE	54.70%	48.21%
S6_A	EGFR	p.L858R	ctDNA	Negative	1.7%
S6_A	EGFR	p.L858R	FFPE	16.77%	16.86%
S9_A	EGFR	p.L858R	ctDNA	Negative	Negative
S9_A	EGFR	p.L858R	FFPE	24.08%	20.20%
S19_A	EGFR	p.L858R	ctDNA	Negative	Negative
S24_A	EGFR	p.L858R	ctDNA	Negative	Negative
S3_A	KRAS	p.G12C	ctDNA	Negative	Negative
S3_A	KRAS	p.G12C	FFPE	24.54%	22.35%

NGS, next generation sequencing; ctDNA, circulating tumour DNA; FFPE, formalin-fixed, paraffin embedded.

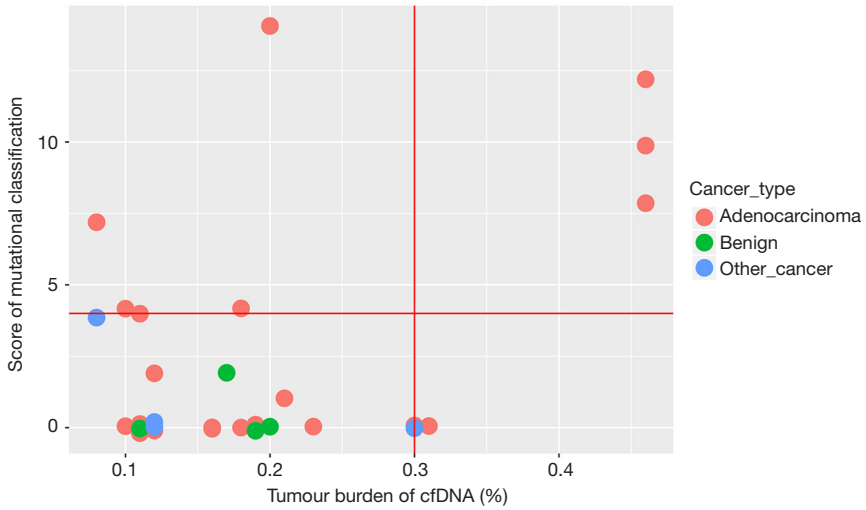

Figure 4 Benign/malignant SPN distribution. TMS and TMB of SPN distribution, which was calculated by in-house software. Different colour indicates different type of SPN. Other cancer, malignant SPN except adenocarcinoma. TMB, tumour mutational burden; TMS, ctDNA mutational score; SPN, solitary pulmonary nodule.

All the potential mutational reads in the panel except germline mutations, which were identified by the normal control, were used to calculate the value:

$$TMD = \frac{sum(N_i)}{sum(D_i)}$$

$i = all\ the\ mutant\ sites\ except\ germline\ mutations$

D_i represents the reads of genomic i-th site; N_i represents the summary reads of non-reference base at potential mutation i.

Based on the method developed in our study, the MS and TMB values of each sample are shown in *Figure 4*. The green dot represents benign samples. Thus, all four benign samples were distributed in the region within TMB ≤0.2, MS ≤2. If TMB =0.3 or MS =4 was used as the cut-off value for malignant SPN prediction, 33.3% of malignant adenocarcinoma samples could be predicted accurately based on the ctDNA samples. In contrast, the sensitivity of CEA (cut-off 10 ng/mL with 97% specificity) was only ~10%, which was lower than the mutation model prediction.

Discussion

LDCT, as an imaging tool for early lung cancer screening, provided insufficient benefit to participants in this study. It is reported that 39.1% of all participants in the LDCT arm of the trial had at least one positive screen, and 96.4% of these initial positive screenings represented false positives for lung cancer (20). Overabundance of false positives could lead to higher screening costs and unnecessary invasive procedures on candidates who do not actually have lung cancer (21). According to our thousands of medical records, we found that nearly 30% of peripheral nodules in lung surgery candidates were non-malignant, and tumour serological markers do not reliably diagnose malignant nodules with high sensitivity. It seemed that protein biomarkers from serum played a less important role and produced false signals during the test. As for the non-malignant cases, some patients underwent operations because of false prediction, whereas most of the rest chose surgery out of fear of the possibility of malignancy. Thus, ctDNA is defined as a more reliable tool to deliver more specific information for both patients' and physicians' reference, also to further define the high-risk population, and to provide a more cost-effective method for diagnosis. ctDNA may provide an opportunity for accurate diagnosis with the advantages of non-invasiveness and no bias of heterogeneity.

Peripheral nodule DNA from surgery candidates had no significant correlation with tumour size and stage, but mutational numbers were significantly different between the benign and malignant nodules. Driver mutations were detected in 71% of malignant nodules. As for DNA mutations from the SPN plasma, advanced tumorigenesis stages and SPN size had no significant influence on somatic

mutations. Moreover, the difference between benign nodules and malignant nodules was not significant. Only ~3.9% of DNA mutations from lung nodules could also be detected in the respective ctDNA by the 10,000-fold sequencing. The concordance of hotspot driver mutations between the malignant nodule DNA and the corresponding ctDNA was only 5.8%, which was much lower than the 85% concordance of cancer tissue DNA and ctDNA in the advanced tumour stage (22). Meanwhile, there was no significant difference in concordance between the stage I and those of stage II and III which were diagnosed postoperatively by incidental finding of positive lymph nodes. This might because the early ctDNA signal of peripheral nodules had not been released into the blood system, or the early DNA mutation frequency was too low to be detected with nowaday sequencing approaches. Thus, improving sensitivity of tumour detection should not be attempted through increasing depth or coverage of sequencing. Somatic mutations were significantly different between benign and malignant tissue DNA but not ctDNA, given that it could not be tested and evaluated with conservative single-gene mutations and hotspot mutations in the early tumour stage due to the possible mechanisms and pathways. Interestingly, somatic mutations were also found in benign nodules, and most of the ctDNA mutations were not detected in FFPE samples, which needed further large-scale validation study.

Mutation concordance (including driver mutation) also suggested that predicting malignant nodules through driver mutation detection based on ctDNA has limited application. This finding encouraged us to grade and score all of the specific mutations to set up a prediction model according to how strongly the mutations correlate with lung cancer. The model first integrated the whole mutational differences, which not only included 'tumour mutational burden' but also evaluated the influence of 'potential mutation'. It overcame the limitation of ctDNA low-frequency mutation detection by NGS. According to this model, we could predict 33.3% of malignant patients (sensitivity) with 100% specificity. Therefore, circulating cfDNA from patients with early lung cancer could reasonably accelerate early diagnosis by ultra-deep sequencing of at least 10,000-folds depth (>1,000-fold unique reads depth) and whole tumour mutation evaluation. This model was the first non-invasive method to predict the malignancy based on ctDNA, which could benefit more than one-third of pulmonary nodule candidates. The potential clinical application of this tool, after extensive validation, is supplemental to LDCT, which yields a great number of false positive cases (7). The high specificity (100%) of the ctDNA genetic model can help us 'rule in' some cases (~30%) that are highly suspected to have malignant disease and should be subjected to surgery with great confidence.

More work shall be done in further studies. Because of the relatively low concordance of tissue DNA and ctDNA mutations, it was obvious that lung cancer genes-targeted capture sequencing was not efficient enough to diagnose with ideal sensitivity by simply increasing sequencing depth or coverage of ctDNA from early candidates. To achieve clinical utility, we propose that sequencing panel contents could be expanded from lung adenocarcinoma to other subtypes to better depict the performance for whole lung nodule patients. This model also shall be optimized by following larger cohort WGS sequencing data and correlated clinical data so that more cancer related gene mutations can be established in this mutational model for more sensitive differentiation in future studies. Therefore, following the remarkable findings of the cfDNA study, ctDNA could still play an important role in diagnosing nodules identified by LDCT or biomarkers as benign or malignant (21). The field is still rushing towards the identification of screening- or diagnostic-specific markers for malignant circulating cfDNA. Other techniques with theoretically higher sensitivity, such as multiplex methylation or cancer-related antibodies detection, might be incorporated to establish a multidimensional, powerful tool for early diagnosis.

Acknowledgements

Funding: This work was supported by National Key R & D Program of China (2016YFC0905400); Guangzhou Key Laboratory of Cancer Trans-Omics Research (GZ2012, NO348) and Guangzhou Science and Technology Project (201400000001-2 and 201400000004-5); Pearl River Nova Program of Guangzhou (No. 201506010065); project 2015B020232008 from Guangdong Province as well as National Precision Medicine Project (SQ2016YFSF090334); Chinese National Natural Science Foundation (Grant No. 81501996); Key Project of Guangzhou Scientific Research Project (Grant No. 201804020030); Guangdong Doctoral Launching Program (Grant No. 2014A030310460); and Doctoral Launching Program of Guangzhou Medical University (Grant No.2014C27); Tianjin Municipal Science and Technology Special Funds for Enterprise Development (No.

14ZXLJSY00320) and special foundation for High-level Talents of Guangdong (2016TX03R171); Key Project of Livelihood Technology of Guangzhou (2011Y2-00024).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The study protocol was reviewed and approved by the Institutional Review Board of the First Affiliated Hospital of Guangzhou Medical University (No. 2015-25). A written informed consent form, describing the purpose of the study, was signed by all of the participants.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin 2016;66:7-30.
2. Rusch VW, Chansky K, Kindler HL, et al. The IASLC Mesothelioma Staging Project: Proposals for the M Descriptors and for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Mesothelioma. J Thorac Oncol 2016;11:2112-9.
3. Kramer BS, Berg CD, Aberle DR, et al. Lung cancer screening with low-dose helical CT: results from the National Lung Screening Trial (NLST). J Med Screen 2011;18:109-11.
4. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 1999;354:99-105.
5. Wahidi MM, Govert JA, Goudar RK, et al. Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007;132:94S-107S.
6. Starnes SL, Reed MF, Meyer CA, et al. Can lung cancer screening by computed tomography be effective in areas with endemic histoplasmosis? J Thorac Cardiovasc Surg 2011;141:688-93.
7. Bartholmai BJ, Koo CW, Johnson GB, et al. Pulmonary nodule characterization, including computer analysis and quantitative features. J Thorac Imaging 2015;30:139-56.
8. Ruilong Z, Daohai X, Li G, et al. Diagnostic value of 18F-FDG-PET/CT for the evaluation of solitary pulmonary nodules: a systematic review and meta-analysis. Nucl Med Commun 2017;38:67-75.
9. Fahrmann JF, Grapov D, DeFelice BC, et al. Serum phosphatidylethanolamine levels distinguish benign from malignant solitary pulmonary nodules and represent a potential diagnostic biomarker for lung cancer. Cancer Biomark 2016;16:609-17.
10. Kupert E, Anderson M, Liu Y, et al. Plasma secretory phospholipase A2-IIa as a potential biomarker for lung cancer in patients with solitary pulmonary nodules. BMC Cancer 2011;11:513.
11. Wang W, Liu M, Wang J, et al. Analysis of the discriminative methods for diagnosis of benign and malignant solitary pulmonary nodules based on serum markers. Oncol Res Treat 2014;37:740-6.
12. Newman AM, Bratman SV, To J, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med 2014;20:548-54.
13. Szpechcinski A, Rudzinski P, Kupis W, et al. Plasma cell-free DNA levels and integrity in patients with chest radiological findings: NSCLC versus benign lung nodules. Cancer Lett 2016;374:202-7.
14. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy for cancer. Clin Chem 2015;61:112-23.
15. Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014;6:224ra24.
16. Izumchenko E, Chang X, Brait M, et al. Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. Nat Commun 2015;6:8258.
17. Li S, Choi YL, Gong Z, et al. Comprehensive Characterization of Oncogenic Drivers in Asian Lung Adenocarcinoma. J Thorac Oncol 2016;11:2129-40. Erratum in: J Thorac Oncol 2017;12:408.
18. Shao D, Lin Y, Liu J, et al. A targeted next-generation sequencing method for identifying clinically relevant mutation profiles in lung adenocarcinoma. Sci Rep 2016;6:22338.
19. Kim EY, Cho EN, Park HS, et al. Compound EGFR mutation is frequently detected with co-mutations of actionable genes and associated with poor clinical outcome in lung adenocarcinoma. Cancer Biol Ther 2016;17:237-45.
20. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395-409.
21. Brothers JF, Hijazi K, Mascaux C, et al. Bridging the clinical gaps: genetic, epigenetic and transcriptomic biomarkers for the early detection of lung cancer in the post-National Lung Screening Trial era. BMC Med

2013;11:168.
22. Lee JY, Qing X, Xiumin W, et al. Longitudinal monitoring of EGFR mutations in plasma predicts outcomes of NSCLC patients treated with EGFR TKIs: Korean Lung Cancer Consortium (KLCC-12-02). Oncotarget 2016;7:6984-93.

Cite this article as: Ye M, Li S, Huang W, Wang C, Liu L, Liu J, Liu J, Pan H, Deng Q, Tang H, Jiang L, Huang W, Chen X, Shao D, Peng Z, Wu R, Zhong J, Wang Z, Zhang X, Kristiansen K, Wang J, Yin Y, Mao M, He J, Liang W. Comprehensive targeted super-deep next generation sequencing enhances differential diagnosis of solitary pulmonary nodules. J Thorac Dis 2018;10(Suppl 7):S820-S829. doi: 10.21037/jtd.2018.04.09

Methods

Patient materials

A total of 1,254 consecutive candidate patients were reviewed following the IRB-approved protocols for resection of lung peripheral nodules in the First Affiliated Hospital of Guangzhou Medical University from January 2015 to November 2016. The 35 plasma and formalin-fixed, paraffin embedded (FFPE) tissue samples were collected from patients with lung peripheral solitary nodules ≤3 cm in diameter of varying size and differentiation. Complete ground glass nodules (GGNs), which were thought to be highly correlated with either non-invasive malignancies or benign changes, were not included in this study. This study is approved by ethical review board of our institution (No. 2015-25).

Blood cell/FFPE cell library preparation and NGS

The library was constructed by shearing peripheral blood cell DNA with an ultrasonoscope to generate fragments with a peak of 250 bps, followed by end repair, A-tailing, and ligation to the Illumina-indexed adapters according to the standard library construction protocol (23). Target enrichment was performed on the designed cancer-related gene capture probe (NimbleGen, Roche Sequencing, Pleasanton, CA, USA). Sequencing was performed with 2×101 bp paired-end reads and an 8-bp index read on an Illumina Hiseq 2,500/4,000 platform (San Diego, CA, USA).

ctDNA library preparation and NGS

Blood samples were collected by different hospitals in China using Cell-Free DNA BCT® blood collection tubes (Streck, La Vista, NE, USA) and transported to a clinical diagnosis lab in Tianjin. The tubes were centrifuged at 1,600 g/min for 10 min. Then, we transferred the plasma to 1.5-mL tubes and centrifuged at 18,000 g/min for 5 min to remove any remaining cells and cellular debris. Finally, we transferred the supernatant to a fresh tube and stored it at –80 ℃. The ctDNA from each 2-mL volume of plasma was extracted using the QIAamp Circulating Nucleic Acid Kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. We quantified the ctDNA isolated from plasma by the Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA, USA). ctDNA purified from plasma was used in the subsequent NGS panel sequencing assays. The library for ctDNA was constructed with the KAPA LTP Library Preparation Kit for Illumina Platforms (Kapa Biosystems, Wilmington, MA, USA) following the manufacturer's instructions without modification (24). Sequencing was performed with 2×101 bp paired-end reads and an 8-bp index read on an Illumina Hiseq 2,500/4,000 platform.

SNV/InDel calling

Raw reads were first processed by removing adaptors and filtering low-quality reads using SOAPnuke (http://soap.genomics.org.cn/) before aligning to the human reference GRCh37 using BWA aligner (v0.6.2-r126) (25) and removing PCR duplications by PICARD (v1.98). Then, local realignment and base quality score recalibration were performed using GATK (v2.3-9) (26). Subsequently, an in-house software was used to call candidate single nucleotide variants (SNV) using the Bayesian model, after which SNV with strand bias and read location bias were filtered using the Fisher's exact test and Kolmogorov-Smirnov test separately (27). Then, SNVs in the local control set were filtered. SNVs were scored according to GC content, adjacent SNV and InDels, multiple mapping locations, and so on. Finally, SNVs with a low score were removed.

Candidate InDels were extracted from the CIGAR information in the BAM files. Next, the de Bruijn method was used to conduct the local *de novo* assemble based on the K-mers from the mapping reads (28). By comparison with the reference sequencing, InDels were predicted. InDels in the corresponding blood cell samples were removed. Finally, InDels in simple repeat regions of the human genome were checked again because of the possibility of more sequencing errors in these regions.

The method for detecting SNV/InDels in the ctDNA samples was the same as that for FFPE sequencing data, except for one additional step that was used to filter the raw mutant set. Twelve-bp paired reads were used as endogenic duplex consensus molecular barcodes and clustered (29). Those with identical barcodes and similar sequences (with consistency >80%) were considered duplication clusters of one template. The order of paired-end sequences was used to identify the sense and anti-sense strands of the template. Only the mutations with both sense and anti-sense strands were used for further analysis.

Somatic SNV and InDels were annotated by ANNOVAR, and only mutations that changed protein structure were retained for further analysis.

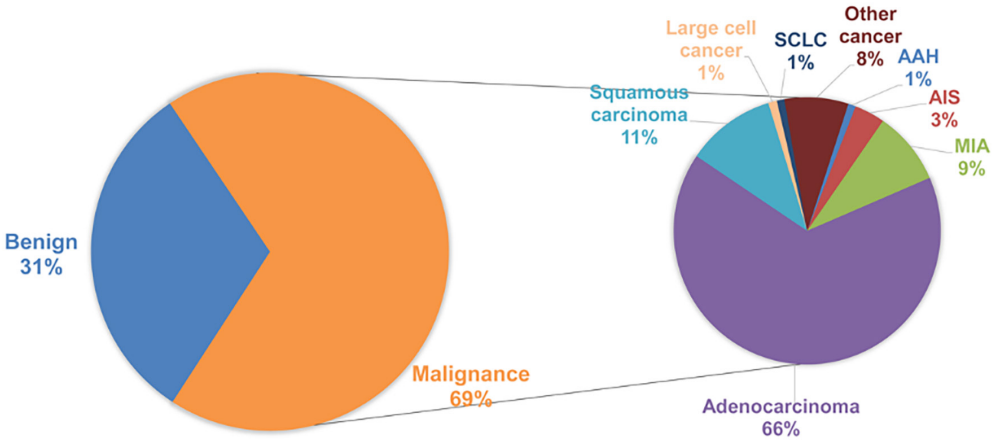

Figure S1 Clinical distribution of all SPNs. SPN, solitary pulmonary nodule; AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; SCLC, small cell lung cancer.

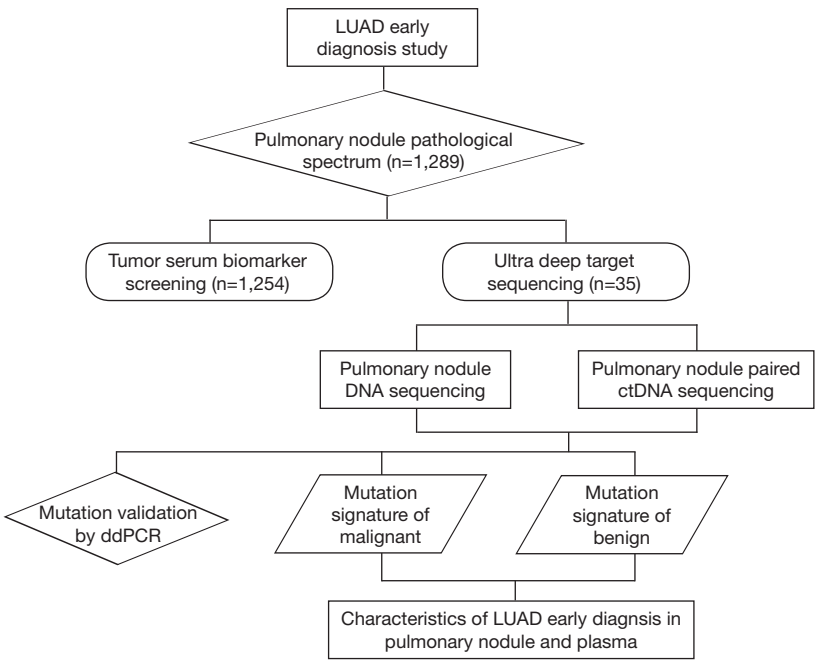

Figure S2 Schedule of pulmonary nodules ultra-deep sequencing and mutation spectrum building. ctDNA, circulating tumour DNA.

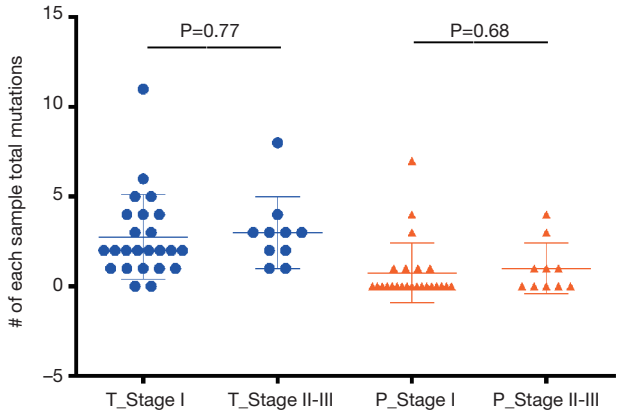

Figure S3 The mutational number comparison between stage I patients and stage II–III patients.

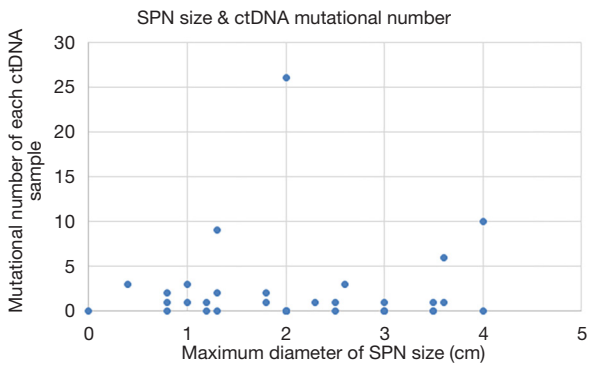

Figure S4 The correlation between the number of ctDNA mutations and the size of SPN, which was measured by the maximum diameter of SPN. SPN, solitary pulmonary nodule; ctDNA, circulating tumour DNA.

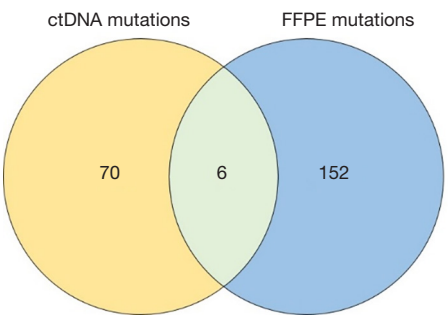

Figure S5 The mutation overlap between the tissue DNA and the corresponding ctDNA sample. FFPE, formalin-fixed, paraffin embedded; ctDNA, circulating tumour DNA.

Table S1 Clinical information for the 38 sequencing samples analyzed in this st

Sample_ID	Gender	Age (year)	Histology	Primary site	Node size (maximum diameter) cm	Node size (D. max) cm	TNM	Stage	CEA (ng/mL)
S32_N	F	62	Benign	Right sided	3.6	3.6×2.8×2.2	–	–	1.24
S33_N	F	53	Benign	Right sided	1.2	1.2×1×1	–	–	1.72
S34_N	M	39	Benign	Left sided	0.4	–	–	–	3.56
S35_N	M	27	Benign	Right sided	3	3×2.5×2	–	–	1.07
S1_A	F	55	Adenocarcinoma	Right sided	2.5	2.5	T1bN2M0	IIIa	1.68
S2_A	M	64	Adenocarcinoma	Left sided	3.5	3.5×3×2.5	T1bN1M0	IIa	1.97
S3_A	M	63	Adenocarcinoma	Left sided	2	2	T1bN0M0	Ib	4.01
S4_A	M	61	Adenocarcinoma	Right sided	2.6	2.6×2×2	T1bN0M0	Ib	3.71
S5_A	M	32	Adenocarcinoma	Unknown	3.5	3.5	T1bN1M0	IIa	1.18
S6_A	F	59	Adenocarcinoma	Right sided	1.8	1.8×1.5×1.5	T1aN1M0	IIa	7.85
S7_A	F	43	Adenocarcinoma	Left sided	1.2	1.2×0.6	T1aN0M0	Ia	0.72
S8_A	M	67	Adenocarcinoma	Right sided	3.6	3.6×3	T2N1M0	IIa	1.8
S9_A	F	47	Adenocarcinoma	Right sided	2.5	2.5×2.5×2.2	T2N0M0	Ib	0.5
S10_A	F	38	Adenocarcinoma	Right sided	4	4×3.5×1.7	T2aNxMx		4.41
S11_A	M	81	Adenocarcinoma	Left sided	2	2×1.5	T1aN0M0	Ia	0.71
S29_O	M	51	Other	Right sided	2.3	2.3×1.3×0.7	T1bN2M0	IIIa	1.88
S12_A	F	70	Adenocarcinoma	Right sided	0.8	1.7×0.5	T1bN0M0	Ib	0.88
S13_A	M	73	Adenocarcinoma	Right sided	2	2×1.5	T2aN0M0	Ib	4.06
S14_A	M	56	Adenocarcinoma	Right sided	1.3	1.3×0.7×0.7	T1aN0M0	Ia	2.29
S15_A	M	45	Adenocarcinoma	Right sided	4	4×3×2.5	T2aN1M0a	Iia	2.23
S16_A	F	43	Adenocarcinoma	Right sided	2	2×1×0.6	T1N0M0	Ia	18.15
S17_A	M	50	Adenocarcinoma	Left sided	NA	2.7×2×1.5	T2aN0M0	Ib	1.44
S18_A	F	60	Adenocarcinoma	Right sided	1	1×1×1	T1aN0M0	Ia	0.55
S28_S	M	69	Squamous cell carcinoma	Left sided	3.5	3.5×3×3	T2aN0M0	Ib	2.2
S19_A	M	52	Adenocarcinoma	Right sided	3	3×2.7	T2aN2M0	IIIa	2.36
S20_A	M	49	Adenocarcinoma	Left sided	3	3×1.8	T1bN0M0	Ib	2.26
S30_O	F	31	Other	Right sided	1.3	1.3×1	T1aN0M0	Ia	0.55
S21_A	F	60	Adenocarcinoma	Left sided	1.3	1.3×0.5	T1aN0M0	Ia	1.41
S22_A	M	70	Adenocarcinoma	Right sided	2	2×1.8×1.5	T1aN0M0	Ia	1.63
S26_A	M	59	AIS	Left sided	0.8	0.8×0.5	T1aN0M0	Ia	2.15
S31_O	M	72	Other	Right sided	2	2×2×1.5	T1aN0M0	Ia	3.69
S27_A	M	64	MIA	Right sided	1	1×0.8	T1aN0M0	Ia	3.34
S23_A	F	49	Adenocarcinoma	Left sided	0.8	–	T1aN0M0	Ia	1.36
S24_A	M	66	Adenocarcinoma	Right sided	1.8	1.8×1.7×1	T1aN0M0	Ia	1.93
S25_A	M	68	Adenocarcinoma	Left sided	3	3×2.5×2	T2aN3M0	IIIb	11.25

AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma.

Table S2 All the gene list in the sequencing panel (27)

OR14C36
PARG
ERBB4
INSRR
DDR2
TMEM199
OR2T33
DSPP
KCNB2
ZNF479
ANAPC1
MB21D2
TSHZ3
MAP1B
THSD4
DNAH8
CNTN5
CDH10
KDR
EPB41L4A
TP53
OR4M2
SNAPC4
NTRK1
PTEN
OR51V1
ZFHX4
KRTAP4-8
NAV3
OR10Z1
PCDH11X
EPHA3
APC
SMAD4
STK11
ZNF804A
DDX11
FGFR4
OR2B11
DNMT3B

Table S2 (*continued*)

ZEB1
CDKN2A
NDUFS1
ADAM23
TBX6
XIRP2
FGFR1
MET
IL32
NTRK3
FAM135B
REG3A
KEAP1
PTPRD
RALGAPB
OR4C16
OR8H2
ATXN1
GAB1
JAK3
REG1B
LRRC56
FGFR2
DCAF4L2
KIAA2022
EPHA5
FGFR3
KIT
CROCC
CNTNAP3B
KRAS
CSNK2A1
INHBA
BRAF
FAM47A
AKT1
JAK2
NOTCH1
PRB2
PDGFRA

Table S2 (*continued*)

PBX2
WDR62
VAV3
CNTNAP2
CHEK2
KIAA0907
NUDT11
RYR2
LRRIQ3
OR4K2
FAM47C
KRTAP5-5
OR2M2
TNRC6A
VGLL3
OR2T34
RET
OR5D18
NF1
RB1
KLK1
FBN2
NRAS
OR4C15
LPA
MMP27
ATXN3
CTNNB1
GNA15
EGFR
ROS1
POTEC
MUC6
FBXW7
CDH12
OR5L2
NYAP2
CLIP1
KRTAP4-11
FOLH1

Table S2 (*continued*)

ZNF804B
PIK3CA
NOTCH2
OR2T2
ALK
AKAP6
NBPF10
NFE2L2
OR10G8
SH2D2A
MUC16
OR4N2
DHX9
ATM
PAPPA2
OR4N4
ERBB2
ZNF814
BEST3
ZNF598

Table S3 Sequencing depth of each sample

Sample_ID	Depth		
	Normal	Tissue	ctDNA
S1_A	761	431	2,948
S10_A	702	636	1,479
S11_A	718	177	3,031
S12_A	765	386	5,643
S13_A	823	1,642	1,896
S14_A	1,428	402	5,889
S15_A	905	220	7,654
S16_A	1,370	275	7,484
S17_A	524	1,911	937
S18_A	801	833	4,454
S19_A	1,597	916	1,024
S2_A	728	197	2,596
S20_A	1,961	1,245	942
S21_A	1,922	601	1,856
S22_A	2,224	459	980
S23_A	1,362	362	805
S24_A	2,032	892	779
S25_A	1,952	818	1,424
S26_A	1,716	1,110	889
S27_A	1,820	1,072	892
S28_S	736	1,772	2,591
S29_O	911	542	3,035
S3_A	757	171	2,606
S30_O	1,803	513	704
S31_O	2,543	1,452	720
S32_N	672	1,089	3,732
S33_N	650	1,441	2,157
S34_N	1,829	1,113	1,593
S35_N	743	516	1,020
S4_A	659	465	3,559
S5_A	619	711	1,890
S6_A	845	268	1,894
S7_A	699	786	2,267
S8_A	660	520	2,275
S9_A	671	409	2,537

Table S4 List of somatic SNV/InDels identified by FFPE samples sequencing data

Sample_ID	Chr	Start_POS	End_POS	Ref	Mutational type	Genotype	Function	Gene	Transcript	Exon	Base mutation	Protein mutation
S1_A	chr7	55259438	55259439	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.2497T>G[:=]	p.[Leu833Val]:[=]
S1_A	chr7	55242468	55242484	TTAAGAGAAGCAACAT	Del	TTAAGAGAAGCAACAT/T	Cds-indel	*EGFR*	NM_005228.3	EX19	c.[2240_2257delTAAGAGAAGCAACAT];[=]	p.[Leu747_Ser752delinsSer];[=]
S1_A	chr10	89717677	89717686	GAAGACAAG	Del	GAAGACAAG/G	Frameshift	*PTEN*	NM_000314.4	EX2	c.[704_711delAAGACAAG];[=]	p.[Glu235_Lys237delinsValfs*?];[=]
S2_A	chr15	22368735	22368734	C	Snv	C/T	Missense	*ORAM2*	NM_001004719.2	EX1E	c.160C>T	p.H54Y | p.His54Tyr
S2_A	chr15	22368881	22368880	C	Snv	C/A	Missense	*ORAM2*	NM_001004719.2	EX1E	c.306C>A	p.F102L | p.Phe102Leu
S2_A	chr17	7577090	7577089	C	Snv	C/G	Missense	*TP53*	NM_000546.5	EX8	c.[848G>C];[=]	p.[Arg283Pro];[=]
S3_A	chr2	168105289	168105291	AG	Del	AG/A	–	*XIRP2*	NM_152381.5	intron	g.[168105291delG];[=]	.
S3_A	chr7	88965976	88965977	C	Snv	C/A	Missense	*ZNF804B*	NM_181646.2	EX4E	c.3681C>A	p.S1227R | p.Ser1227Arg
S3_A	chr11	123901085	123901086	C	Snv	C/A	Missense	*OR10G8*	NM_001004464.1	–	c.757C>A	p.P253T | p.Pro253Thr
S3_A	chr12	25398284	25398285	C	Snv	C/A	Missense	*KRAS*	NM_004985.3	EX2	c.[34G>T];[=]	p.[Gly12Cys];[=]
S3_A	chr14	20296030	20296031	T	Snv	T/G	Missense	*OR4N2*	NM_001004723.1	–	c.424T>G	p.Y142D | p.Tyr142Asp
S3_A	chr14	20296033	20296034	G	Snv	G/A	Missense	*OR4N2*	NM_001004723.1	–	c.427G>A	p.A143T | p.Ala143Thr
S3_A	chr17	7578405	7578406	C	Snv	C/T	Missense	*TP53*	NM_000546.5	EX5	c.[524G>A];[=]	p.[Arg175His];[=]
S3_A	chr17	29559189	29559190	A	Snv	A/G	Coding-synon	*NF1*	NM_000267.3	EX25	c.[329?A>G];[=]	p.[=];[=]
S3_A	chr19	907341 7	907341 8	G	Snv	G/T	Coding-synon	*MUC16*	–	–	c.14028C>A	.
S3_A	chr19	58382263	58382264	T	Snv	T/C	Utr-3	*ZNF814*	–	–	g.[58382264T>C];[=]	.
S4_A	chr1	156834221	156834222	T	Ins	T/TT	Splice-5	*NTRK1*	NM_001007792.1	IVS3	c.[197+2_197+3insT];[=]	.
S4_A	chr1	162725076	162725077	C	Snv	C/T	Coding-synon	*DDR2*	NM_001014796.1	EX7	c.[549C>T];[=]	p.[=];[=]
S4_A	chr7	55242464	55242480	GGAATTAAGAGAAGCA	Del	GGAATTAAGAGAAGCA/G	Cds-indel	*EGFR*	NM_005228.3	EX19	c.[2236_2250delGAATTAAGAGAAGCA];[=]	p.[Glu746_Ala750delfs*?];[=]
S4_A	chr7	55242464	55242480	TG	Del	TG/T	Frameshift	*TP53*	NM_000546.5	EX9	c.[991del[C];[=	p.[Gln331Argfs*?];[=]
S4_A	chr17	7579493	7579494	T	Snv	T/A	Nonsense	*TP53*	NM_000546.5	EX4	c.[193A>T];[=]	p.[Arg65*];[=]
S4_A	chr19	9060496	9060497	T	Snv	T/C	Missense	*MUC16*	–	–	c.26949A>G	p.Ile898Met | p.Ile898Met
S5_A	chr15	88690675	88690676	G	Snv	G/A	Intron	*NTRK3*	NM_001007156.2	IVS5	c.[396-42C>T];[=]	.
S6_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.[2573T>G];[=]	p.[Leu858Arg];[=]
S6_A	chr15	88524582	88524578	TTTG	Del	TTTG/T	Cds-indel	*NTRK3*	NM_001007156.2	EX16	c.[1595_1597delCAA];[=]	p.[Ser532_Asn533delinsTyr];[=]
S6_A	chrX	91133541	91133542	T	Snv	T/A	Missense	*PCDH11X*	–	–	c.2303T>A	p.V768D | p.Val768Asp
S7_A	chr7	55242477	55242478	G	Snv	G/C	Missense	*EGFR*	NM_005228.3	EX19	c.[2248G>C];[=]	p.[Ala750Pro];[=]
S7_A	chr7	55242467	55242477	ATTAAGAGAA	Del	ATTAAGAGAA/A	Cds-indel	*EGFR*	NM_005228.3	EX19	c.[2239_2247delTTAAGAGAA];[=]	p.[Leu747_Glu749delfs*?];[=]
S8_A	chr1	248616493	248616492	G	Snv	G/T	–	–	–	–	g.[248616493G>T];[=]	.
S8_A	chr2	168103079	168103080	C	Snv	C/T	–	–	–	–	g.[168103080C>T];[=]	.
S8_A	chr7	55249004	55249005	G	Snv	G/T	Missense	*EGFR*	NM_005228.3	EX20	c.2303G>T];[=]	p.[Ser768Ile];[=]
S8_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.2573T>G];[=]	p.[Leu858Arg];[=]
S8_A	chr10	43610858	43610859	G	Snv	G/C	Intron	*RET*	NM_020630.4	IVS11	c.[2136+67SG>C];[=]	.
S8_A	chr15	88429074	88429076	CA	Del	CA/C	Intron	*NTRK3*	NM_001012338.2	IVS17	c.[2134-110delT];[=]	.
S8_A	chr17	75781901	75781902	A	Ins	A/ATA	Frameshift	*TP53*	NM_000546.5	EX6	c.[657_658insTA];[=]	p.[Tyr220fs*?];[=]
S9_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.2573T>G];[=]	p.[Leu858Arg];[=]
S9_A	chr17	7578180	7578181	G	Snv	G/C	Missense	*TP53*	NM_000546.5	EX6	c.[668C>G];[=]	p.[Pro223Arg];[=]
S9_A	chr17	7579866	7579867	G	Snv	G/A	Nonsense	*TP53*	NM_000546.5	EX2	c.[46C>T];[=]	p.[Gln16*];[=]
S9_A	chrX	73960816	73960817	G	Snv	G/A	Missense	*KIAA2022*	–	–	c.[3575C>T];[=]	p.[Ser1192Phe];[=]
S9_A	chr7	55242469	55242470	T	Snv	T/C	Missense	*EGFR*	NM_005228.3	EX19	c.[2240T>C];[=]	p.[Leu747Ser];[=]
S10_A	chr17	7577538	7577539	G	Snv	G/A	Missense	*TP53*	NM_000546.5	EX7	c.[742C>T];[=]	p.[Arg248Trp];[=]
S11_A	chr5	71495573	71495574	C	Snv	C/T	Missense	*MAP1B*	–	–	c.[6392C>T];[=]	p.[Pro2131Leu];[=]
S11_A	chr8	139164634	139164635	C	Snv	C/T	Missense	*FAM135B*	–	–	c.[2083G>A];[=]	p.[Val695Ile];[=]
S11_A	chr10	123256191	123256192	G	Snv	G/A	Nonsense	*FGFR2*	NM_000141.4	EX13	c.[1717C>T];[=]	p.[Arg573*];[=]
S11_A	chr7	55242463	55242479	AGGAATTAAGAGAAGC	Del	AGGAATTAAGAGAAGC/A	Cds-indel	*EGFR*	NM_005228.3	EX19	c.[2235_2249delGAATTAAGAGAAGC];[=]	p.[Lys745_Ala750delinsLys];[=]
S11_A	chr7	17656407	17656407 6	G	Snv	G/A	Missense	*PAPPA2*	NM_020318.2	EX3	c.[1337G>A];[=]	p.[Ser446Asn];[=]
S12_A	chr1	156843655	156843656	T	Snv	T/C	Missense	*NTRK1*	NM_001007792.1	EX9	c.[992T>C];[=]	p.[Leu331Pro];[=]
S12_A	chr1	248512497	248512498	A	Snv	A/G	Missense	*OR14C36*	NM_001001918.1	EX1E	c.[422A>G];[=]	p.[Gln141Arg];[=]
S12_A	chr4	55139766	55139767	C	Snv	C/A	Coding-synon	*PDGFRA*	NM_006206.4	EX10	c.[1428C>A];[=]	p.[=];[=]
S12_A	chr6	16328689	16328690	C	Snv	C/T	Utr-5	*ATXN1*	NM_000332.3	EX8	c.[-149G>A];[=]	.
S12_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.[2573T>G];[=]	p.[Leu858Arg];[=]
S12_A	chr8	139207677	139207678	C	Snv	C/A	Intron	*FAM135B*	NM_015912.3	IVS8	c.[824-128G>T];[=]	.
S12_A	chr15	88423539	88423540	G	Snv	G/A	Coding-synon	*NTRK3*	NM_001012338.2	EX19	c.[2295C>T];[=]	p.[=];[=]
S12_A	chr17	29575816	29575817	A	Snv	A/G	Intron	*NF1*	NM_000267.3	IVS29	c.[3975-185A>G];[=]	.
S13_A	chr1	74574911	74574912	A	Snv	A/G	Intron	*LRRIQ3*	NM_001105659.1	IVS5	c.[867+166T>C];[=]	.
S13_A	chr2	168100349	168100348	C	Snv	C/T	Missense	*XIRP2*	NM_001199144.1	EX7	c.[1781C>T];[=]	p.[Pro594Leu];[=]
S13_A	chr3	178951724	178951725	G	Snv	G/A	Intron	*PIK3CA*	NM_006218.2	IVS20	c.[2937-157G>A];[=]	.
S13_A	chr4	55139556	55139557	C	Snv	C/T	Intron	*PDGFRA*	NM_006206.4	IVS9	c.[1365-147C>T];[=]	.
S13_A	chr4	55976706	55976707	G	Snv	G/A	Missense	*KDR*	NM_002253.2	EX9	c.1118C>T];[=]	p.[Ser373Phe];[=]
S13_A	chr8	77618390	77618391	G	Snv	G/A	Missense	*ZFHX4*	NM_024721.4	EX2	c.2068G>A];[=]	p.[Glu690Lys];[=]
S13_A	chr8	139164914	139164915	G	Snv	G/T	Missense	*FAM135B*	NM_015912.3	EX13	c.1803C>A];[=]	p.[His601Gln];[=]
S13_A	chr9	83412151	83412160	G	Snv	G/T	Missense	*PTPRD*	NM_001040712.2	EX25	c.3770C>A];[=]	p.[Pro1257Gln];[=]
S13_A	chr9	84863619	84863620	G	Snv	G/T	Intron	*PTPRD*	NM_001040712.2	IVS11	c.[1814-1038C>A];[=]	.
S13_A	chr17	7577123	7577124	C	Snv	C/A	Missense	*TP53*	NM_000546.5	EX8	c.[814G>T];[=]	p.[Val272Leu];[=]
S13_A	chr17	29553464	29553465	G	Snv	G/A	Missense	*NF1*	NM_000267.3	EX18	c.2014G>A];[=]	p.[Gly672Arg];[=]
S13_A	chr17	29592382	29592383	C	Snv	C/T	Intron	*NF1*	NM_000267.3	IVS35	c.[4772+26C>T];[=]	.
S13_A	chr19	1218272	1218273	C	Snv	C/T	Intron	*STK11*	NM_000455.4	IVS1	c.[291-143C>T];[=]	.
S13_A	chr19	31767735	31767736	G	Snv	G/T	Missense	*TSHZ3*	NM_020856.2	EX2E	c.[2963C>A];[=]	p.[Pro988His];[=]
S13_A	chr19	31767736	31767737	G	Snv	G/T	Missense	*TSHZ3*	NM_020856.2	EX2E	c.[2962C>A];[=]	p.[Pro988Thr];[=]
S13_A	chr19	31770028	31770029	C	Snv	C/T	Missense	*TSHZ3*	NM_020856.2	EX2E	c.[67OG>A];[=]	p.[Asp224Asn];[=]
S14_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.2573T>G];[=]	p.[Leu858Arg];[=]
S14_A	chr7	55260497	55260498	A	Snv	A/G	Missense	*EGFR*	NM_005228.3	EX22	c.[2665A>G];[=]	p.[Arg889Gly];[=]
S15_A	chr6	117622322	117622323	G	Snv	G/A	Intron	*ROS1*	NM_002944.2	IVS41	c.[6570-23C>T];[=]	.
S15_A	chr7	140453135	140453136	A	Snv	A/T	Missense	*BRAF*	NM_004333.4	EX15	c.[1799T>A];[=]	p.[Val600Glu];[=]
S16_A	chr17	7577533	7577534	C	Snv	C/A	Missense	*TP53*	NM_000546.5	EX7	c.[747G>T];[=]	p.[Arg249Ser];[=]
S16_A	chr7	55242463	55242479	AGGAATTAAGAGAAGC	Del	AGGAATTAAGAGAAGC/A	Cds-indel	*EGFR*	NM_005228.3	EX19	c.[2235_2249delGAATTAAGAGAAGC];[=]	p.[Lys745_Ala750delinsLys];[=]
S17_A	chr1	248511912	248511913	T	Snv	T/C	–	–	–	–	g.[248511913T>C];[=]	.
S17_A	chr2	79313418	79313419	C	Snv	C/T	Intron	*REG1B*	NM_006507.3	IVS4	c.[321+74G>A];[=]	.
S17_A	chr15	88726716	88726717	G	Snv	G/C	Coding-synon	*NTRK3*	NM_001007156.2	EX5	c.327C>G];[=]	p.[=];[=]
S17_A	chr17	7579306	7579307	C	Snv	C/A	Intron	*TP53*	NM_000546.5	IVS4	c.[375+5G>T];[=]	.
S17_A	chr19	9060523	9060524	G	Snv	G/T	Coding-synon	*MUC16*	NM_024690.2	EX3	c.[26922C>A];[=]	p.[=];[=]
S17_A	chrX	34149098	34149099	C	Snv	C/A	Missense	*FAM47A*	NM_203408.3	EX1E	c.[1297G>T];[=]	p.[Asp433Tyr];[=]
S17_A	chr7	55259523	55259524	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.[2582T>G];[=]	p.[Leu861Arg];[=]
S18_A	chr17	7577558	7577559	G	Snv	G/A	Missense	*TP53*	NM_000546.5	EX7	c.[722C>T];[=]	p.[Ser241Phe];[=]
S18_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.[2573T>G];[=]	p.[Leu858Arg];[=]
S18_A	chrX	91133541	91133542	T	Snv	T/A	Missense	*PCDH11X*	–	–	c.2303T>A	p.V768D | p.Val768Asp
S19_A	chr10	89720787	89720789	GG	Del	GG/G	Frameshift	*PTEN*	NM_000314.4	EX8	c.[940del[G];[=	p.[Glu314Lysfs*?];[=]
S19_A	chr18	48584584	48584586	TT	Del	TT/T	Frameshift	*SMAD4*	NM_005359.5	EX6	c.[759del[T];[=	p.[Phe253Leufs*?];[=]
S19_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.[2573T>G];[=]	p.[Leu858Arg];[=]
S19_A	chr7	55242468	55242487	TTAAGAGAAGCAACATCTC	Del	TTAAGAGAAGCAACATCTC/T	Cds-indel	*EGFR*	NM_005228.3	EX19	c.[2240_2257delTAAGAGAAGCAACATCTC];[=]	p.[Leu747_Pro753delinsSer];[=]
S19_A	chr3	89448596	89448597	C	Snv	C/A	Missense	*EPHA3*	NM_005233.5	EX7	c.[1560C>A];[=]	p.[Asn520Lys];[=]
S19_A	chr19	12184813	12184814	A	Snv	A/G	Splice-3	*STK11*	NM_000455.4	IVS1	c.[291-2A>G];[=]	.
S25_A	chr19	9060744	9060746	TT	Del	TT/T	–	–	–	–	g.[9060746delT];[=]	.
S25_A	chr19	9060744	9060746	AGCCCTCCAGGTCCCCA	Del	AGCCCTCCAGGTCCCCA/A	Intron	*TP53*	NM_000546.5	IVS6	c.[96+20_96+35delTGGGGACCTGAGGGC];[96+20_96+35delTGGGGACCTGAGGGC]	.
S25_A	chr7	55242468	55242478	G	Snv	G/A	Missense	*EGFR*	NM_005228.3	EX19	c.[2248G>C];[=]	p.[Ala750Pro];[=]
S25_A	chr7	55242477	55242478	ATTAAGAGAA	Del	ATTAAGAGAA/A	Cds-indel	*EGFR*	NM_005228.3	EX19	c.[2239_2247delTTAAGAGAA];[=]	p.[Leu747_Glu749delfs*?];[=]
S25_A	chr9	85048493	85048492	A	Snv	A/G	Intron	*PTPRD*	NM_002839.3	IVS10	c.[1669-33A>G];[=]	.
S25_A	chr17	75782578	75782577	C	Snv	C/G	Missense	*TP53*	NM_000267.3	EX30	c.[3986C>G];[=]	p.[Ser1329*];[=]
S25_A	chr2	17809895	17809896	A	Snv	A/G	Missense	*NFE2L2*	NM_001145412.2	EX2	c.[411>C];[=]	p.[Leu14Pro];[=]
S25_A	chr2	17809895	17809896	A	Snv	A/G	Missense	*NFE2L2*	NM_006164.4	EX2	c.[89T>C];[=]	p.[Leu30Pro];[=]
S25_A	chr6	90569888	90569889	C	Snv	C/G	–	–	–	–	g.[90569889C>G];[=]	.
S25_A	chr9	90586823	90586824	G	Snv	G/A	–	–	–	–	g.[90586824G>A];[=]	.
S25_A	chr19	10597049	10597050	G	Snv	G/A	Utr-3	*KEAP1*	NM_012289.3	EX6E	c.*278C>T];[=]	.
S25_A	chr19	24761526 1	24761526 3	AA	Del	AA/A	–	–	–	–	g.[24761526 3delA];[=]	.
S25_A	chr17	39253415	39253420	AGACA	Del	AGACA/A	–	–	–	–	g.[39253417_39253420delGACA];[=]	.
S25_A	chr1	162724923	162724924	T	Snv	T/A	Intron	*DDR2*	NM_001014796.1	IVS6	c.418-22T>A];[=]	.
S25_A	chr5	112174255	112174256	G	Snv	G/A	Missense	*APC*	NM_000038.5	EX16E	c.2965G>A];[=]	p.[Asp989Asn];[=]
S25_A	chr7	55259514	55259515	T	Snv	T/G	Missense	*EGFR*	NM_005228.3	EX21	c.2573T>G];[=]	p.[Leu858Arg];[=]
S25_A	chr7	116412077	116412078	T	Snv	T/C	Intron	*MET*	NM_001127500.1	IVS14	c.[3082+35T>C];[=]	.
S25_A	chr7	116412077	116412078	T	Snv	T/C	Intron	*MET*	NM_000245.2	IVS14	c.[3028+35T>C];[=]	.
S25_A	chr9	8492980	8492981	T	Snv	T/C	Splice-3	*PTPRD*	NM_002839.3	IVS26	c.[2350-2A>G];[=]	.
S25_A	chr1	156844882	156844883	G	Snv	G/T	Intron	*NTRK1*	NM_001007792.1	IVS11	c.[1246+83G>T];[=]	.
S25_A	chr1	247615150	247615151	C	Snv	C/A	–	–	–	–	g.[247615151C>A];[=]	.
S25_A	chr1	248616179	248616180	C	Snv	C/A	–	–	–	–	g.[248616180C>A];[=]	.
S25_A	chr1	248737759	248737760	G	Snv	G/T	–	–	–	–	g.[248737760G>T];[=]	.
S25_A	chr4	55960837	55960838	G	Snv	G/T	Intron	*KDR*	NM_002253.2	IVS21	c.[2971+131C>A];[=]	.
S25_A	chr7	57187926	57187927	G	Snv	G/T	–	–	–	–	g.[57187927G>T];[=]	.
S25_A	chr8	88886060	88886061	G	Snv	G/T	–	–	–	–	g.[88886061G>T];[=]	.
S25_A	chr11	49168536	49168537	T	Snv	T/C	–	–	–	–	g.[49168537T>C];[=]	.
S25_A	chr16	32890555	32890556	C	Snv	C/A	–	–	–	–	g.[32890556C>A];[=]	.
S25_A	chr17	7578202	7578203	C	Snv	C/G	Missense	*TP53*	NM_000546.5	EX6	c.[646G>T];[=]	p.[Val216Leu];[=]
S25_A	chr17	7578202	7578203	C	Snv	C/A	Missense	*TP53*	NM_001276760.1	EX6	c.[529G>T];[=]	p.[Val177Leu];[=]
S25_A	chr17	7578204	7578205	C	Snv	C/A	Missense	*TP53*	NM_000546.5	EX6	c.[644G>T];[=]	p.[Ser215Ile];[=]
S25_A	chr19	12184813	12184814	A	Snv	A/G	Splice-3	*STK11*	NM_000455.4	IVS1	c.[291-2A>G];[=]	.
S25_A	chr19	9060744	9060746	TT	Del	TT/T	–	–	–	–	g.[9060746delT];[=]	.

Table S5 List of structure variation genes detected in FFPE samples

Sample_ID	Gender	Age	Histology	Gene1	Gene2	Cancer_SoftClip_Sup	Normal_SoftClip_Sup	All_freq (%)	Validation by IHC
S5_A	M	32	Adenocarcinoma	ALK	EML4	22	0	4.14	Yes
S27_A	M	64	MIA	KIF5B	RET	46	0	5.79	Yes
S23_A	F	49	Adenocarcinoma	ALK	EML4	29	0	4.65	Yes
S23_A	F	49	Adenocarcinoma	ERC1	ROS1	22	0	4.92	No

FFPE, formalin-fixed, paraffin embedded; MIA, minimally invasive adenocarcinoma.

Table S6 List of somatic SNV/InDels identified in ctDNA samples

Sample_ID	Chr	Start_POS	End_POS	Ref	Mutational type	Genotype	Function	Gene	Transcript	Exon	Base mutation	Protein mutation
S4_A	chr7	55266511	55266512	A	Snv	A/T	Missense	EGFR	NM_005228.3	EX23	c.[2804A>T];[=]	p.[Gln935Leu];[=]
S8_A	chr7	55249004	55249005	G	Snv	G/T	Missense	EGFR	NM_005228.3	EX20	c.[2303G>T];[=]	p.[Ser768Ile];[=]
S8_A	chr7	55259514	55259515	T	Snv	T/G	Missense	EGFR	NM_005228.3	EX21	c.[2573T>G];[=]	p.[Leu858Arg];[=]
S8_A	chr10	43610858	43610859	G	Snv	G/C	Intron	RET	NM_020630.4	IVS11	c.[2136+675G>C];[=]	–
S14_A	chr10	43610448	43610449	G	Snv	G/T	Intron	RET	NM_020630.4	IVS11	c.[2136+265G>T];[=]	–
S15_A	chr7	55269080	55269081	G	Snv	G/T	Intron	EGFR	NM_005228.3	IVS25	c.[3114+33G>T];[=]	–
S15_A	chr10	43610677	43610678	G	Snv	G/T	Intron	RET	NM_020630.4	IVS11	c.[2136+494G>T];[=]	–
S15_A	chr10	43611827	43611828	T	Snv	T/A	Intron	RET	NM_020630.4	IVS11	c.[2137-204T>A];[=]	–
S27_A	chr7	55219055	55219056	G	Snv	G/A	Splice-5	EGFR	NM_005228.3	IVS5	c.[628+1G>A];[=]	–
S25_A	chr6	117708978	117708979	A	Snv	A/C	Missense	ROS1	NM_002944.2	EX13	c.[1978T>G];[=]	p.[Trp660Gly];[=]

SV calling

Chimeric read pairs were collected and clustered to detect structural variations (SVs). The clipped parts of the soft clipped reads were collected and mapped to the genome (30). Genome locations of clipped and remaining parts were clustered to determine the accurate break points of SVs.

ddPCR

Droplet digital PCR (ddPCR, QuantStudio 3D Digital PCR System, Life Technologies, Carlsbad, CA, USA) was performed in this study. According to the guidebooks, QuantStudio 3D Digital PCR Master Mix v2 and TaqMan Assay were thawed to room temperature and mixed approximately 10 times. The targeted DNA was diluted to 200–2,000 copies/μL. The reaction mixture was prepared following the recommended protocol, and then the mixture was loaded into the QuantStudio 3D Digital PCR Chip as soon as possible.

References

23. Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques 2014;56:61-4, 66, 68, passim.
24. Xu S, Lou F, Wu Y, et al. Circulating tumor DNA identified by targeted sequencing in advanced-stage non-small cell lung cancer patients. Cancer Lett 2016;370:324-31.
25. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589-95.
26. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.
27. Kan Z, Zheng H, Liu X, et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res 2013;23:1422-33.
28. Narzisi G, O'Rawe JA, Iossifov I, et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nat Methods 2014;11:1033-6.
29. Lanman RB, Mortimer SA, Zill OA, et al. Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA. PLoS One 2015;10:e0140712.
30. Wang J, Mullighan CG, Easton J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 2011;8:652-4.